International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Peculiar Approach for Detection of Cluster Outlier in High Dimensional Data

Abhimanyu Kumar

Suresh Gyan Vihar University, Jagatpura, Jaipur, Rajasthan

Abstract: ***As data is becoming huge and available in diverse formats, we need algorithms enabling data to be clustered and detecting the outliers. We have many methods of dealing with the outlier values Outliers are those unprecedented values that don't go with the structure of the data set and hence lead to wrong results or confusion. Data mining algorithms work by first detecting the outlier values and then either filling these values or removes these values. After thorough search, we found that outliers and clusters have a connection between them hence its very useful to deal with both these concepts for efficient data analysis. So in this paper we introduce an algorithm which is based on k means [1] for the detection of clusters and outliers that aim to detect the clusters and the outliers. Using this algorithm we have made clusters based on the relation between cluster and the inter dependence between the cluster and the outliers, running until termination is achieved***

Keywords: K means, Clusters, Outliers, Data analysis

1. Introduction

As data is becoming huge and available in diverse formats, we need algorithms enabling data to be clustered and detecting the outliers, in recent times we have had many methods focusing on clustering and outlier detection[2,3,4,5].As data sets in today's time are huge and in diversified format, hence leading to inconsistencies, redundancies. Thus reducing the effectiveness of the algorithm during our study of different data sets we found out that clusters and outliers have some connection between them. So it is important to deal clusters and outliers as concepts of the same significance in data analysis. The biggest worry for the data analyst is the detection of outliers and clusters on the main attributes of the data sets and therefore results are based on these concepts Hard Reality is that the results of the data sets and actual results don't match, therefore we have focused on using a methodology which makes clusters not only based upon the data value types but also on dependence between the clusters and outliers.

2. Problem Descriptions

Both these concepts are interlinked to each other. In reality, data sets are very complex and thus having diversified format and clustering is a very challenging task. Following section explains the detection procedure Firstly We has used many notations which are defined as follows Let *N* denote the total number of data points and *D* denote the data space dimensionality. Let the input dataset be: B

B= {B1,B2,.....,BN}

Which is normalized to be in the hyper cube $[0,1]^d \subset R^d$. Now each data point is a dimensional vector.

$$\overrightarrow{A_i} = [A_{i1}, A_{i2}, \ldots\ldots, A_{id}].$$

For the given input parameters we run our algorithm recursively. Initially we assume the number of clusters to be Ac and the current numbers of outliers is Ao. The set of clusters are $C'=\{C',C',\ldots,C'_{Ac}\}$,and the set of outliers $O'=\{O', O', \ldots\ldots, O'_{a0}\}$. We define compactness to measure the quality of cluster based upon the closeness with the centroid.

2.1 Cluster Compactness

A cluster consists of the data points who have a close relation to each other rather than outside points, From the literature [6,7], relation between cluster is evaluated using separation ,while relation within the clusters is evaluated using compactness. Given set of clusters $C'=\{C',C',\ldots,C'_{Fc}\}$and the set of outliers $O'=\{O',O',\ldots\ldots,O'_{Fo}\}$,the compactness (CP)is evaluated using the following formula

$$CP(Ci) = \frac{\sum_{p \in C_i} d(p, {}^{k}C_i)}{|C_i|}$$

Where ${}^{k}C_i$ is the centroid of the Cluster C_i, p is a data point in Cluster C_i, $|C_i|$ number of data points in C_i, and $d(p, {}^{k}C_i)$ is the distance between p and ${}^{k}C_i$. The centroid ${}^{k}C_i$ of the cluster is the algebraic average of all the points in the cluster:

$${}^{k}C_i = \frac{\sum_{p \in C_i} p}{|C_i|}$$

2.2 Diversities of Data Groups

Diversity in general is used to measure or the differences between any two concepts, here also its used to describe the difference between the two clusters, the difference between the two outliers and the one difference between the cluster and an outlier. The procedure used for measuring the diversities is based on measuring the distance between the points. It would be preferable to use the compactness

Methodology instead of the diversity approach,
Diversity between a cluster c and an outlier O is:

$$D_1(c, O) = w_1.d_{min}(O,c) + w_2.d_{avr}(O,c)$$

Where

$$W_1 = \frac{1}{CP(C)+1}, \quad W_2 \frac{\pm CP(C)}{CP(C)+1}, \quad d_{avr}(O,c) = d(O, K_c)$$

dmin (O, c)=max(d(O,Kc)–rmax,0) where rmax is the data points distance in C from its centroid. The procedure for assigning up the weights W1 and W2 are similar as in[8].We simply combine both the concepts to get a parameters based on which we can measure the closeness between the points within a cluster. Diversity between two clusters $C_1$and $C_2$is:

$$D_2(C_1,C_2)=\frac{d(C_1,C_2)}{cp\ (c_1)+\ cp\ (c_2)}$$

Where d (C_1,C_2) be the average distance between the two clusters or the minimum distance between them. Diversity between two outliers o_1 and o_2 is:

$$d_3\,(o_1,o_2)= d\,(o_1,o_2)$$

2.3Qualities of Data Groups

For any cluster to be of highest quality, we need to measure the diversity both between the clusters but also from the outliers. Let c be close to one outlier then its quality will be affected by the presence of the outlier as the outlier is meant to be away from the cluster.
Quality of the cluster c is:

$$q(c)=\frac{\frac{\sum_{c'\in C,c'\neq c} D_2(c,c')}{F_c-1}+\frac{\sum_{o\in O} D_1(c,o)}{F_o}}{Cp(c)}$$

Whenever the larger Q (c) is, the quality of the cluster c will be better.

Similarly the quality of an outlier o is:

$$q\,(o)=\frac{\sum_{o'\in O,o'\neq o} D_2(o,o')}{F_o-1}+\frac{\sum_{c\in C} D_1(c,o)}{F_c}$$

The larger q(o) is, the outlier o will attain the better quality.

3. Cluster Outlier Detection

The primary task of this procedure is to generate set of cluster and outliers from a given set of input data. There is a direct correspondence between the clusters and the outliers. What our algorithm does is that it creates cluster on relationship between the clusters as well as within the cluster and the outliers. Our approach is a greedy approach similar to[9].The process, runs until termination is achieved

Functionality of the algorithm is based on two stages. In the first stage, the main task is to find and identify the cluster points and also finding the outliers. The second stage is very crucial as in this stage these data sets are further refined. The process of refining is as follows, we calculate the distance of all points from all the cluster center and then compare these distances and place them in the cluster for which the distance is minimum.

3.1 First Stage

For this stage the most important task is the identification of the medoids or the centroid. After selecting these points, the main task is to assemble the data points as different subsets with these medoids being the centers. The placement of these points is based on criteria as follows, we calculate the distance between the point and all the medoids and place the point in the set for which the distance is minimum, after performing these steps we need to determine if these are outliers or not. In the following section, we have expailned all the methods.

3.1.1Acquiring Medoids

Finding out correct medoids is very crucial for our algorithm. We have used a similar approach as mentioned in [10]. We choose at random the set of data points and the number of data points required are equal to the number of cluster needed. Then we use the approach defined in [9] on all these sets (rs1 and rs2). After applying the Greedy approach [9], the efficiency is improved.

3.1.2 Dispatch the Data Points

After creating the initial clusters, we need to examine if the Data points within the cluster are closer to each other and further away from points in the other clusters This step is done by finding center of all the created clusters and redoing the initial process until at consecutive stages we are getting the same cluster.

3.1.3 Dataset Division

Now as we get the set h of the initial division of the input. dataset B, now we have to check the size of each medoid associated with the subset of the data H . Now we exploit some kind of method which adjusts the criterion for determining that whether a medoid is an outlier or it belongs to the cluster. After the completion of the process of cluster or outlier, it should be contrary that the size of the cluster set c < k if the initial sizes RS1 and RS2 are large enough. If it happens again we just run the initial setup to make sure that the size of the cluster set c is at least k.

3.2 Second Stage

At this juncture we combine all the clusters to create K cluster. While combining the clusters we check the quality ro seagrigate the outlier values. Quality of the cluster is determined using the relationship within the clusters and the relationship between the clusters and outliers. After this, the third step will constitute of methods for identifying the boundary values leading to outliers. And in the fourth step we will refine the set of clusters and outliers continuously by optimally exchanging the selected boundary data points and the worst qualities of outliers. These steps are repeated until certain termination is reached.

3.2.1 Merging

Antacedently exploiting the outlier detection process we have to first amalgamate the current set of cluster C' to K cluster. This operation is an iterative one in each iteration phase, whichever two nearest or closest clusters are found in C' they are amalgamated together. According to the diversity measurement D2 (C1, C2) of two clusters (defined in the part 2.2) the distance between the clusters C1 and C2 is calculated. The iteration step is performed perpetually until the total number of clusters in C' is K. Now we have to compute the centroid of each cluster Ci C' (denoted as ci).

3.2.2 Grouping of Clusters and Outliers

Now for each outlier O we have to find its nearest cluster C. According to the diversity measurement D1 (C, O) which is (defined in part 2.2) the distance between the cluster C and outlier O is calculated. And according to the information of the outliers in O and the information of the clusters in C the quality Q (O) (which is defined in the part 2.3) is calculated. The worst qualities of outliers are put into the set O'. Thus, for each cluster C C. According to not only the information of the clusters in C but also the information of outliers in O the quality Q (C) (defined in the part 2.3) is calculated. The worst qualities of clusters are put into set C'.

3.2.3Findboundarydatapoints

We need those data points into the clusters that are not only more distant from the centroid of the clusters but also contain the smallest number of neighboring data points as the data points of the clusters. The latest circumstances assure that this method not only promote clusters of standard geometries such as hyper spherical.

3.2.4Exchangeofdatapoints

In this step it is use to exchange the outliers and the boundary data points characteristics. Now for each outlier O in we add it into its closest cluster. For each boundary data point bp in BP we change it into a new outlier. Thus, we do not exchange the data points of the boundary between clusters are that whole data division quality will be degraded if it is carried.

Algorithm (K: Number of clusters)
Start

1. First stage
Repeat
A and B are the two equivalent constant to K, where A>B;
RS1 = A. K;
RS2 = B. K;
T1 = it is the random set having the size of RS1;
T2 = Finding K medoids from T1 having size of RS2;
H Dispatch the data points (T2);
C and O it is the cluster or the outlier ();
Until ≥ K

2. Second stage
C Merging of the Cluster (C);
Repeat
For each outlier o O do
Start
Find the nearest cluster C
Stop

Group the current set of clusters and current set of outliers in ascending order according to their qualities; Now exchange the cluster and the outlier (); O' it is the set of outliers according to the worst qualities; BP it is the set of boundary data points according to the worst qualities;
U = O' BP;
Until (U gives a constant value or the iteration ≥ Ω)
Stop.

3.3 Analysis of Time and Space

Let us assume size of data set is n. Now during the whole process we need to keep track of information on all points which collectively holds O (n) space. For the second stage or we can say the iteration stage we need space for the information of current set of clusters C and the current set of outliers O, the boundary data points for each cluster, the worst qualities of outliers and clusters in each iteration. The total amount of space needed is O (n). The time required for each iteration is

$$O\ (n + |C| log|C| + |O| log|O|)$$

particularly for the computation of the various types of qualities and sorting process. C and O. So the total amount of time required for the algorithm is O $(\Omega * (n + |C| log|C| + |O| log|O|))$ in which Ω is known as the threshold for the number of iterations.

3.4 Workflow of the Algorithm

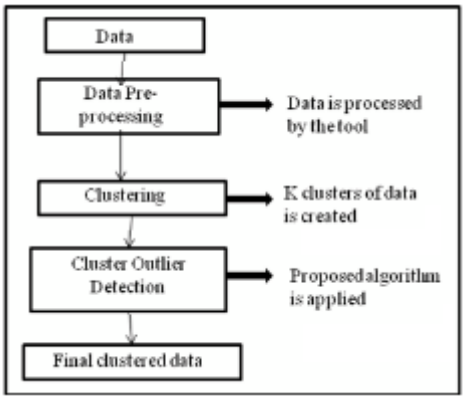

Figure1: Work flow of the algorithm

Above is the figure depicting the workflow chart of our proposed algorithm? Firstly the data set is taken and feed into the tool for pre-processing of the data and after the completion of this step K clusters of Data is being created. Then our proposed algorithm is applied on the data for cluster outlier detection and in the last we get the final clustered data.

4. Experiments and Results

In this section we have done the experiment on Ecoli data set which is taken from [11]. We conducted our experiment on Cluster 3.0 tool [12]. We use an Ecoli data set which is used to find Localization site of proteins. The dataset contains a total of 336 instances (objects) each having attributes (1 name and 7 input features). It contains eight clusters having sizes of (143, 77, 52, 35, 20, 5, 2 and 2). Now we perform our algorithm on the Ecoli data and after the experiment we observe that our proposed algorithm has the most information about the first three largest clusters.

Table 1: Clustering result of the algorithm for Ecoli data

	K=0	K=1	K=2	K=3	K=4	K=5	K=6	K=7
C (R)	143	77	52	35	20	5	2	2
C (D)	138	80	73	N	N	4	N	N
C(R)∩C(D)	132	54	50	N	N	4	N	N
Accuracy%	95.6	67.50	68.49	N	N	100	N	N
Recall Value %	92.3	70.12	96.15	N	N	80	N	N

Now from the result table we see that the data set contains 8 clusters C^* (R) for K = (0 to 7). There are three clusters which are too small so we set the clusters parameter number K to 5. And the accuracy of the detected cluster measured according to the accuracy % and the recall value %. Now for C^* (D) detected cluster and C^*(R) for the real cluster we calculate the accuracy of C^* (D) with respect to C^* (R) as $\frac{C^*(D) \cap C^*(R)}{C^*(D)}$ and recal value is $\frac{C^*(D) \cap C^*(R)}{C^*(R)}$. Hence C^* (D) is called as comparable cluster of C^*(R) if the accuracy and the recall value of C^*(D) and C^*(R) are high.

5. Conclusions

In this paper we interpolate a peculiar approach for cluster outlier detection in high dimensional data. This approach can be able to ameliorate the clusters and outliers qualities for those high dimensional data which contains noise. The clusters are ascertained and managed according to the intra-relationship within the clusters and inter-relationship between the clusters and the outliers. The whole management and modification of the clusters and outliers are done repeatedly just before a certain termination is reached. Now further dealing with the clusters and outliers as a concept of the same significance in data analysis. We also concern about the flushing the difficulties of the deficiency of match between the ground truth of the real data and their obtainable characteristics.

References

[1] J. MacQueen. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability. Volume I, Statistics. 1967.
[2] D. Yu, G. Sheikholeslami and A. Zhang. Findout: Finding outliers in very large datasets. The Knowledge and Information System, (4), October 2000.
[3] Charu C. Aggarwal et al. Outlier Detection for high dimensional data. In SIGMOID Conference, 2001
[4] M.F. Jiang et al. Two phase clustering process for outlier detection. Pattern Recognition Letters 22 pages 691-700. 2001
[5] Angiuli et al. Fast outlier detection in high dimensional spaces. In the proceedings of KDD. 2002.
[6] Chi-Farn Chen, Jyh-Ming Lee. The Validity Measurement of Fuzzy C-Means Classifier for Remotely Sensed Images. In Proc. ACRS 2001 - 22nd Asian Conference on Remote Sensing, 2001.
[7] Maria Halkidi et al. A data set oriented Approach for clustering Algorithm Selection. In PKDD, 2001.
[8] Dantong Yu and Aidong Zhang. ClusterTree: Integration of Cluster Representation and Nearest Neighbor Search for Large Datasets with High Dimensionality. IEEE Transactions on Knowledge and Data Engineering (TKDE), 14(3), May/June 2003.
[9] T. Gonzalez. Clustering to minimize the maximum intercluster distance. Theoretical Computer Science, 38:311-322, 1985.
[10] Charu C. Aggarwal et al. Fast algorithms for projected clustering. In the proceedings of the ACM SIGMOID Conference on management of Data, pages 61-72, 1999
[11] The UCI KDD Archive [http://Kdd.ics.uci.edu]. University of California, Irvine, Department of Information and Computer Science.
[12] Michiel de Hoon. [http://bonsai.ims.u-tokyo.ac.jp], Open source clustering software. Institute of medical Science University of Tokyo.

Author Profile

Abhimanyu Kumar was born in India in 1991. He completed his B. Tech degree with distinction in Computer Science from Suresh Gyan Vihar University, Jaipur, India and now working in **Xerox Corporation** as a Business analyst in Big Data Management Team. His research interest includes A Peculiar approach for cluster outlier detection in high dimensional data.